Statistical Science
2016, Vol. 31, No. 1, 27–29
DOI: 10.1214/15-STS537
Main article DOI: 10.1214/15-STS529

Exploiting the Feller Coupling for the Ewens Sampling Formula

Richard Arratia, A. D. Barbour and Simon Tavaré

We congratulate Harry Crane on a masterful survey, showing the universal character of the Ewens sampling formula.

There are two grand ways to get a simple handle on the Ewens sampling formula; one is the Chinese restaurant coupling, and the other is the Feller coupling. Since Crane has discussed the Chinese Restaurant process, but not the Feller coupling, we will give a brief survey of the latter.

The Ewens sampling formula, given in Crane's (1), has an interpretation in terms of the cycle type of a random permutation of n objects. For $\theta = 1$, it is just Cauchy's formula, expressed in terms of the *fraction* of permutations of n objects that have exactly m_i cycles of order i, $1 \le i \le n$. For general θ, the power

$$\theta^{m_1+m_2+\cdots+m_n} = \theta^K$$

appearing in the formula, where K denotes the number of cycles, biases the uniform random choice of a permutation by weighting with the factor θ^K, the remaining factors involving θ merely reflecting the new normalization constant required to specify a probability distribution. We use the notation $(C_1(n), \dots, C_n(n))$ to denote a random object distributed according to the Ewens sampling formula, suppressing the parameter θ but making explicit the parameter n, so that, with Crane's notation (1),

$$\begin{aligned}(1)\qquad &\mathbb{P}(C_1(n) = m_1, \dots, C_n(n) = m_n)\\ &\quad = p(m_1, \dots, m_n; \theta).\end{aligned}$$

Richard Arratia is Professor, Department of Mathematics, University of Southern California, 3620 S. Vermont Ave, KAP 104, Los Angeles, California 90089-2532, USA (e-mail: rarratia@usc.edu). A. D. Barbour is Professor Emeritus, Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland (e-mail: a.d.barbour@math.uzh.ch). Simon Tavaré is Professor, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, United Kingdom (e-mail: st321@cam.ac.uk).

The Feller coupling, motivated by the example in Feller ([6], page 815) is defined as follows. Take independent Bernoulli random variables ξ_i, $i = 1, 2, 3, \dots,$ with the simple odds ratios $\mathbb{P}(\xi_i = 0)/\mathbb{P}(\xi_i = 1) = (i-1)/\theta$. Thus, $\mathbb{E}\xi_i = \mathbb{P}(\xi_i = 1) = \theta/(\theta + i - 1)$, and $\mathbb{P}(\xi_i = 0) = (i-1)/(\theta + i - 1)$. Say that an ℓ-spacing occurs in a sequence $a_1, a_2, \dots,$ of zeros and ones, starting at position $i - \ell$ and ending at position i, if $a_{i-\ell}a_{i-\ell+1}\cdots a_{i-1}a_i = 10^{\ell-1}1$, a one followed by $\ell - 1$ zeros followed by another one. Then if, for each $\ell \ge 1$, we define

$$C_\ell(n) := \text{the number of } \ell\text{-spacings in}$$
$$\xi_1, \xi_2, \dots, \xi_{n-1}, \xi_n, 1, 0, 0, \dots,$$

the joint distribution of $C_1(n), \dots, C_n(n)$ *is* the Ewens sampling formula, as per Crane's (1) and our (1). This can be seen directly, for the case $\theta = 1$: consider a random permutation of 1 to n, write the canonical cycle notation one symbol at a time, and let ξ_i indicate the decision to complete a cycle, when there is an i-way choice of which element to assign next. The general case $\theta > 0$ follows by biasing, with respect to θ^K: since $K = \xi_1 + \cdots + \xi_n$, and the $\xi_1, \dots, \xi_n$ are independent, biasing their joint distribution by $\theta^{\xi_1+\cdots+\xi_n} = \theta^{\xi_1}\cdots\theta^{\xi_n}$ preserves their independence and Bernoulli distributions, while changing the odds $\mathbb{P}(\xi_i = 0)/\mathbb{P}(\xi_i = 1)$ from $(i-1)/1$ to $(i-1)/\theta$.

Now, the wonderful thing that happens is that, with Y_ℓ *defined* to be the number of ℓ-spacings in the infinite sequence $\xi_1, \xi_2, \dots,$ it turns out that $Y_1, Y_2, \dots$ are mutually independent, and that Y_ℓ is Poisson distributed, with $\mathbb{E}Y_\ell = \theta/\ell$, as in formula (11) in Section 3.8. This shows that the Ewens sampling formula is closely related to the simpler independent process $Y_1, Y_2, \dots, Y_n$. Explicitly, let R_n be the position of the rightmost one in $\xi_1, \xi_2, \dots, \xi_{n-1}, \xi_n$—noting that always $\xi_1 = 1$ so R_n is well-defined—and let $J_n := (n+1) - R_n$. We have

$$(2)\qquad C_\ell(n) \le Y_\ell + 1(J_n = \ell), \qquad 1 \le \ell \le n,$$

with contributions to strict inequality whenever, for some $1 \le \ell \le n$, an ℓ-spacing occurred in $\xi_1, \xi_2, \dots$ starting at $i - \ell$ and ending at $i > n$.

We view (2) as saying that the Ewens sampling formula distributed $(C_1(n), \dots, C_n(n))$ can be constructed from the independent Poisson Y's using at most one insertion, together with a random number of deletions. The expected number of deletions is $O_\theta(1)$, that is, bounded over all n, with the upper bound depending on the value of θ. A concrete upper bound is given in [3], but the limit value, call it $c(\theta)$, is cleaner. This limit *is* the expected number of spacings of length at most 1, with right end greater than 1, in the scale invariant Poisson process on $(0, \infty)$ with intensity $\theta/x\,dx$; see [1]. We have

$$\begin{aligned} c(\theta) &= \int_{x>1} \frac{\theta}{x}\mathbb{P}\bigl(\text{at least one arrival in } (x-1, x)\bigr)\,dx \\ &= \int_{x>1} \frac{\theta}{x}\left(1 - \exp\left(-\int_{x-1}^{x} \frac{\theta}{y}\,dy\right)\right)dx \\ &= \int_{x>1} \frac{\theta}{x}\left(1 - \left(1 - \frac{1}{x}\right)^{\theta}\right)dx \end{aligned}$$

and, using the substitution $v = 1 - 1/x$, we get

$$\begin{aligned} c(\theta) &= \theta \int_0^1 (1-v)^{-1}(1 - v^{\theta})\,dv \\ &= \theta \sum_{n\ge 0}\left(\frac{1}{n+1} - \frac{1}{n+1+\theta}\right) \\ &= \theta\left(\frac{1}{\theta} + \sum_{n\ge 0}\left(\frac{1}{n+1} - \frac{1}{n+\theta}\right)\right) \\ &= 1 + \theta\bigl(\gamma + \psi(\theta)\bigr), \end{aligned}$$

where γ is Euler's constant and ψ is the digamma function.

The simple fact that one can transform the Ewens sampling formula into the highly tractable Poisson process $Y_1, Y_2, \dots, Y_n$ using a bounded (in expectation) number of insertions and deletions is, in itself, quite powerful, since there are interesting aspects of the joint distribution which are insensitive to a bounded number of insertions and deletions. For example, consider the Erdős–Turán law for the order of a random permutation. The order of a permutation is the least common multiple of the lengths of its cycles, and the Erdős–Turán law is the statement of convergence to the standard normal distribution, for the log of the order, centered by subtracting an asymptotic mean $\log^2 n/2$, and scaling by dividing by an asymptotic standard deviation, $\log^{3/2} n/3$. The effect of a finite number of cycle lengths is washed away by the scaling; see [5] for details.

In a similar spirit, and modeled after the Feller coupling for the Ewens sampling formula, [2] shows that for a random integer chosen uniformly from 1 to n, the counts $C_p(n)$ of prime factors, including multiplicity, can be coupled to independent $Z_2, Z_3, Z_5, \dots$ with $\mathbb{P}(Z_p \ge k) = p^{-k}$ for prime p and $k = 0, 1, 2, \dots$ in such a way that $\mathbb{E}\sum_{p\le n} |C_p(n) - Z_p| \le 2 + O((\log\log n)^2/\log n)$; informally, the prime factorization can be converted into the process of independent geometric random variables, using on average no more than $2 + \varepsilon_n$ insertions and deletions. The fact of being able to convert with $o(\log\log n)$ insertions and deletions already easily implies the Hardy–Ramanujan theorem for the normal order of the number of prime divisors, and the fact of being able to convert with $o(\sqrt{\log\log n})$ insertions and deletions readily implies the Erdős–Kac central limit theorem for the number of prime divisors.

The Feller coupling expresses the Ewens sampling formula in terms of the spacings of the independent Bernoulli sequence $\xi_1, \xi_2, \dots, \xi_n$. The conditioning relation, described in Crane's article at the start of Section 3.8, expresses the Ewens sampling formula in terms of the independent Poisson $Y_1, Y_2, \dots, Y_n$. Both these independent processes have *the same* limit upon rescaling, namely, the scale invariant Poisson process on $(0, \infty)$ with intensity $\theta/x\,dx$. This leads to a property of the scale invariant Poisson process: the set of its spacings has *the same* distribution as the set of its arrivals. This property can be exploited to bound the distance to the Poisson–Dirichlet limit, which is mentioned in Crane's Section 4.2. Write $(X_1, X_2, \dots)$ for the random vector distributed according to the Poisson–Dirichlet$(\theta = 1)$. For random permutations, writing $L_i(n)$ for the size of the ith largest cycle, [4] shows that there are couplings which achieve $\mathbb{E}\sum_{i\ge 1} |L_i(n) - nX_i| \sim \frac{1}{4}\log n$, and that no coupling can achieve a constant smaller than $1/4$. For prime factorizations, writing $P_i(n)$ for the ith largest prime factor of a random integer distributed uniformly from 1 to n, [2] shows that there is a coupling of random integers to Poisson–Dirichlet such that $\mathbb{E}\sum_{i\ge 1} |\log P_i(n) - (\log n)X_i| = O(\log\log n)$, and the conjecture that $O(1)$ can be achieved remains open.

REFERENCES

[1] ARRATIA, R. (1998). On the central role of scale invariant Poisson processes on $(0, \infty)$. In *Microsurveys in Discrete Probability* (*Princeton*, *NJ*, 1997). *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **41** 21–41. Amer. Math. Soc., Providence, RI. MR1630407

[2] ARRATIA, R. (2002). On the amount of dependence in the prime factorization of a uniform random integer. In *Contemporary Combinatorics*. *Bolyai Soc*. *Math*. *Stud*. **10** 29–91. János Bolyai Math. Soc., Budapest. MR1919568

[3] ARRATIA, R., BARBOUR, A. D. and TAVARÉ, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann*. *Appl*. *Probab.* **2** 519–535. MR1177897

[4] ARRATIA, R., BARBOUR, A. D. and TAVARÉ, S. (2006). A tale of three couplings: Poisson–Dirichlet and GEM approximations for random permutations. *Combin*. *Probab*. *Comput*. **15** 31–62. MR2195574

[5] BARBOUR, A. D. and TAVARÉ, S. (1994). A rate for the Erdős–Turán law. *Combin*. *Probab*. *Comput*. **3** 167–176. MR1288438

[6] FELLER, W. (1945). The fundamental limit theorems in probability. *Bull*. *Amer*. *Math*. *Soc*. (*N*.*S*.) **51** 800–832. MR0013252